# zoo.csv に関する R 言語を使用した決定木構築

文理学部情報科学科
5419045 高林 秀

2021 年 6 月 27 日

**概要**

本稿では、今年度データ科学 2 で学習した「決定木構築手法」を使用して、本学部ページにて配布されたデータである zoo.csv の決定木構築を実験するものである。また、決定木構築に際し、分割基準や木の高さなどのパラメータをいくつか変更しながら実験を行う。また得られた決定木の評価指標値として、精度の算出を行い、木の良し悪しを判定する。

## 1　目的

本稿では実際に、R 言語を使用し配布データである zoo.csv の決定木構築を行うことで、本年度データ科学 2 で学習した決定木構築の手法への理解を深め、その定着を図ることを目的とする。また、1 年次に学習した latex を用いた PDF 作成の復習も兼ねるものとする。

## 2　理論説明

今回の実験で用いた、計算理論をそれぞれ説明する。

### 2.1　分類学習について

我々の生きている世界には様々なデータが存在する。例えば、農業を行う際「豊作になる条件」を知るには肥料の種類や量、光量や日照時間、温度、雨量等の様々な「属性の値」をもとに予測することが可能である。これを予測する手法として機械学習が挙げられるだろう。機械学習では、コンピュータがある問題とその答えを使用して学習を行い、データに潜むパターン等を識別、発見する技術である。この機械学習は大きく 3 つに系統が別れている。初めに「教師あり学習」、次に「教師なし学習」、最後に「強化学習」である。そしてそこから更に、求める結果や手法によって「分類」「回帰」「クラスタリング」「次元削減」「Q-Learning」と細かく分割される。中でも、今回扱う「分類」はデータが属するクラスを予測することを目的とする。予測するクラスが 2 つならば「2 値分類」、それ以上ならば「多クラス分類」と呼ばれる。

詳細な説明は本稿では行わないが、今回扱う「決定木」は教師あり学習の分類に属する手法である。

これまで、分類の手法として「k-近傍法」等を学習してきた。k-近傍法は、最近傍のデータを k 個選択し、それらが最も多く属するクラスに識別、分類を行う方法であった。決定木では、k-近傍法の手法とは異なり「木構造」を利用する。ある属性の属性値によってデータを徐々に分割していきクラス分類をする。決定木の詳細については後述する「決定木 (Decision tree)」の部分で説明する。

なお分類学習には、ここで紹介した手法以外に「サポートベクターマシン (Support Vector Machine:通称 SVM)」や 「ナイーブベイズ (Naive Bayes)」などがある。最近ではこれら従来の手法に加え、ニューラルネットワークを利用した手法が画像分類等の分野で広がっている (CNN など)。

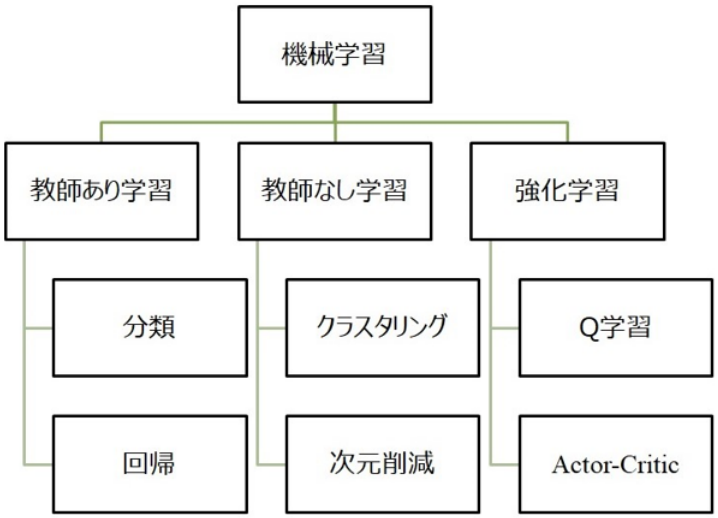


図 1　機械学習の枠組み

## 2.2　決定木 (Decision tree)

決定木とは「木構造を利用した機械学習手法」である。分類を行う決定木を「分類木」、回帰 (連続値の予測) を行う決定木は「回帰木」と呼ばれる。決定木では任意の属性の属性値による条件分岐によって、データを徐々に分割することで結果を出力する。したがって、生成される木構造の枝は分割の結果ラベルを、葉は予測、分類されるクラスの結果を、各ノードは属性に関する分割テスト含んでいる。

決定木を使用する例として、ミカンとリンゴを分類する場合を考える。下記の図のように、ミカンの画像を 4 枚、リンゴの画像を 2 枚の計 6 枚の画像データセットがあるとする。これを、ある条件 A を定め、それに当てはまるもの、そうでないものを分割する。この操作を分割テストという。分割テストの結果によって次の分割テスト行うか否かが決定され、最終的に、ミカンとリンゴが図のように分類される。これが決定木の大まかな流れである。

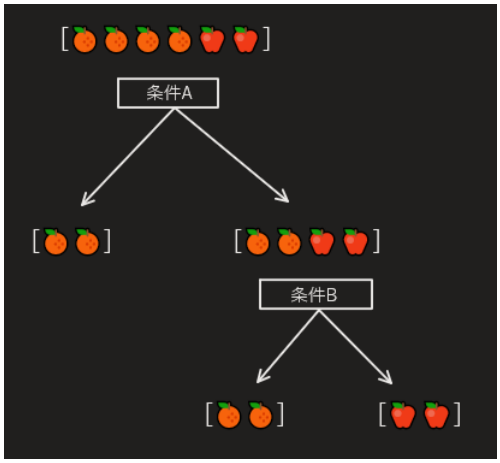


図 2　ミカンとリンゴの分類木

### 2.2.1　決定木のメリット・デメリット

決定木には、前項で述べた SVM 等の他の機械学習手法よりも**分類課程が明確である**というメリットが存在する。これは、分類の結果の理由が他の手法よりも明確である点が挙げられるだろう。例えば先程のミカンとリンゴの例のように条件を満たすか否かによって、データを分割し、ミカンかリンゴか分類する。このとき何故ミカン (またはリンゴ) と分類したのかの理由が「条件 A(または条件 B) を満たしているから」と容易に説明できるであろう。

このように、機械学習を行った結果として生成されるモデルにおいて、判断の仕組みが分かる、説明できるものをホワイトボックス、反対に説明できないあるいは説明しづらいものをブラックボックスと呼ぶ。またホワイトボックスであるようなモデルを備えた人工知能を「説明可能な AI[*1](XAI : Explainable AI)」と呼ぶ。XAI は米国国防高等研究計画局 (DARPA) が手動する研究プロジェクトが発端となった。

以下に、決定木のメリット・デメリットをまとめた表を示す。

表 1　決定木のメリット・デメリット

| メリット | デメリット |
|---|---|
| ・学習結果の可読性が高く結果の根拠を説明しやすい<br>・データの前処理が少なく済むことが多い<br>・予測時に必要な計算量が小さい<br>・回帰、分類の両方に対応可能 | ・条件分岐が複雑になるほど過学習しやすい<br>・精度が突出して良いわけではない |

なお、下記サイトにおいて説明可能な AI の詳細な記載があるのでそのリンクを示す。

`https://blog.global.fujitsu.com/jp/2018-12-27/01/`

このように決定木は、分類課程が分かりやすいためエキスパートシステム[*2]等に利用されることが多い。

[*1] 説明可能な AI : このように、モデルの解釈性をもたらす研究が近年注目されており、ブラックボックス型のモデルを解釈する手法として「SHAP」「LIME」などがあり、Python パッケージが公開されるなど実務での利用が進んでいる。

[*2] エキスパートシステム：ある分野における専門知識を蓄積し、その分野の専門家のように振る舞うことができる。

決定木は、分割テストに使用する属性によって分割結果が変わることは、容易に想像することができる。データの属するクラスの割合をクラス分布と呼ぶ。決定木の分割テストでは、クラスを分類したいのでデータのクラス分布が偏れば偏るほど良い。しかし、分割テストに使用する属性によってクラス分布が偏ったり、均等になってしまいクラス分布があまり変化しない、といったことが発生する。そこで、分割テストに使用する属性を決定するためにいくつか評価基準が存在する。

## 2.3 分割に用いる属性の選択基準

分割テストに使用される属性は、下記のような評価基準を例に選出される。

### 2.3.1 情報利得

情報利得とは一言で言えば「クラスの偏りがどの程度進んだか」を表す数値である。データセット $D$ のおける属性 $A$ の情報利得の計算式は下記。

$$Gain_A(D) = H(D) - H_A(D)$$

この値が大きいほど、分割テストに適した良い属性ということになる。

**■情報量とエントロピー**

上記式の意味を説明するにあたって抑えておかなければならない概念がある。自己情報量 (選択情報量) とエントロピー (平均情報量) である。

情報の数量的構造に関して甘利氏の著書「情報理論」[1] に以下の記述があるので引用する。

> 情報の数量的構造を論ずるにあたって、まず情報とは何であるかを考えなくてはなるまい。「(中略)」、すべての情報に共通な本質を抽象しよう。それは、「情報とはわれわれに何事かを教えてくれるものであり、われわれの不確実な知識を確実にしてくれるものである」というあたりまえのことである。「(中略)」。情報の量はその情報をもらったことによって知識の不確実さがどのくらい減ったかで計れば良いからである。

すなわち、情報量とは、情報を得る前からその情報を得た後の差分で定義するということだ。自己情報量とは下記式で表される数値である。

$$-\log_2 p[bit]$$

これは、確率 $p$ の事象が発生したことを知らせる情報に含まれている情報量を示している。これが導かれるまでの課程は、甘利氏の著書「情報理論」[1] の第 1 章「情報の数量的認識」の部分に記載があるのでそちらを参照いただきたい。この自己情報量は確率 $p$ が大きければ小さい値を取るという性質がある。言い換えれば、当たり前の事象が起こっても受け取れる情報は小さい、ということであり、これは、めったに起きない事象のほうが受け取る情報量が多いということを示している。

次に、エントロピーの説明をする。エントロピーとは平均情報量とも呼ばれる。これも、甘利氏の著書「情報理論」[1] に記載があるので引用する。

前項では、確率 $p$ の事象 $A$ が起こったときは、この情報量は $-\log_2 p$ であることを論じた。「(中略)」。話をもう少し一般的にして $A_1, A_2, ..., A_n$ の $n$ 個の事象があって、それぞれ $p_1, p_2, ..., p_n$ の確率で生ずる場合を考えよう。

$$\Sigma_{i=1}^{n} p_i = 1$$

である。ここで、どの事象が起こったかを教えてもらうことにする。得られる情報の量は、どの $A$ が生じたかで異なってくる。すなわち、$A_1$ が起これば $-\log_2 p_1$、$A_2$ ならば $-\log_2 p_2, ...$ という情報が得られる。「(中略)」、得られる情報の量の期待値は $-\log_2 p_i$ を確率 $p_i$ で平均したもの

$$I = -\Sigma_{i=1}^{n} p_i \log_2 p_i$$

である。

これの $I$ がエントロピーである。すなわち、自己情報量の平均 (期待値) を表している。言い換えれば、どの $A_i$ が起こったかを聞くときに得られる情報量である。また、エントロピーについて甘利氏の著書「情報理論」[1] では次のように記されている。

> われわれが情報をほしいのは、不確定な状況を確定したいからである。この場合、どういう情報がもらえるかは事前にわかるはずがなく、したがって、もらえる情報量そのものはわからない。わかるのはもらえる情報量の期待値だけである。この値は、不確定な状況を確定するのに要する平均情報量だといってもよい。

すなわち、エントロピーとは「情報の不確定さ、程度」を表す数値である、ということだ。エントロピーの性質に関しては本稿では取り扱わないが、甘利氏の著書「情報理論」[1] の第 1 章 20 ページに記載があるので、それを参照いただきたい。エントロピーは、クラス分布のばらつきが大きい時にエントロピーは値が小さくなり、分布が均等であるときには最大値 1 を取る性質がある。

ここまでの話を、先程の情報利得の話に当てはめてみる。このときの自己情報量を $I(c, D)|c$ : クラス, $D$ : データセット (データ集合) とすると、

$$I(c, D) = -\log_2 P_D(c)$$

※ $P_D(c)$ : $D$ 中のデータのクラスが $c$ となる確率

これは言い換えれば、確率 $P_D(c)$ で $D$ のデータのクラスが $c$ と分類されるときの情報量と捉えることができる。

そして、$D$ を分割する前のエントロピーが $H(D)$ で、属性 $A$ での分割後のエントロピーが $H_A(D)$ であり、それぞれ下記式で示すことができる。

$$H(D) = \Sigma_{c \in C} P_D(c) \times I(c, D) = -\Sigma_{c \in C} P_D(c) \log_2 P_D(c)$$

$$H_A(D) = \Sigma_{a \in A} p_D(a) \times H(D_a)$$

$P_D(a)$：$D$ 中のデータの属性 $A$ の属性値が $a$ となる確率。

すなわち、クラス分布を偏らせるためには、情報利得が大きい属性を選択すれば良い。ここまでの話をまとめると、**情報利得 ＝ 分割前のエントロピー - 分割後の各エントロピーの重み付き平均**ということになる。

### 2.3.2 情報利得比

情報利得には、分割数の大きい属性に対して、不当に高い値を返す問題がある。例として、ID 番号 (データの順番号) を分割属性に使用したとき、すべてのデータを別々の部分集合に分割することができるが、機械学習の目的である予測をするという意味において、全く役に立たない。そこで、分割数による正規化の必要性が生じる。そこで、新たな属性選択の評価基準として情報利得が挙げられる。

情報利得比とは、情報利得を分割情報量で正規化した数値である。

**■分割情報量**　分割数が大きい属性に対して、より大きな値をとる性質を持つ。

$$SI_A(D) = \Sigma_{a \in A} P_D(a) \times I(a, D) = \Sigma_{a \in A} P_D(a) \log_2 P_D(a)$$

情報利得比 $GainRaito_A(D)$ は下記式で計算される。

$$GainRaito_A(D) = \frac{Gain_A(D)}{SI_A(D)}$$

### 2.3.3 ジニ係数 (Gini Index)

分割後のデータ集合の、クラス分布が偏っている時、次の 2 つのことが言える。

- その集合から取り出した 2 つの事例が同一のクラスに属する確率が高い。
- その集合から取り出した 2 つの事例が同一のクラスでない確率が低い。

この状況を利用した属性の評価基準がジニ係数である。ジニ係数では「2 つの事例が同一のクラスでない確率」について考える。

前にも述べたように、決定木では、同一のクラスに属さない確率が低い方がクラスが偏っていて良いとされる。ここで、属性 $A$ に対するジニ係数を以下のように定める。

$$Gini_A(D) = \Sigma_{a \in A} P_D(a) \times G(D_a)$$
$$G(D) = 1 - \Sigma_{c \in C} P_D(c)^2$$

$P_D(c)^2$: 2 事例があるクラス $c$ に属する確率。

なお、ジニ係数は、集合から取り出した 2 つの事例が同一のクラスでない確率が低い、ということを表す数値なので、情報利得とは異なり、**より小さい値**の属性がよいとされる。

### 2.4 決定木構築手法

決定木の構築にあたって考慮しなければならないことがある。それはどのような木が望ましいかどうか、である。その際に以下 2 つの基準が存在する。

- 木のコンパクトさ、単純さ
- 木の正確さ

すなわち望ましい木とは、「コンパクトかつ正確なもの」と言えるだろう。ここまで述べたように、決定木は分割に使用する属性によって木の構造が大きく変わる性質がある。したがって、最終的に形成される木の組み合わせは、非常に膨大な数になる。そのため、考えられるすべての木を計算し、評価するのには限界がある。よって、比較的計算量が少なく、簡単な方法で望ましい決定木を作る必要が生じる。この、簡単な方法で望ましい決定木を作る方法を「ヒューリスティックな方法」と呼ぶ。

決定木構築法の代表例として TDIDT(Top Down Induction of Decision Trees) と呼ばれる手法が存在する。

#### 2.4.1 TDIDT

### 2.5 枝刈りと汎化性能 (ロバスト性)

### 2.6 決定木の評価指標

## 3 計算機実験

## 4 まとめ

## 参考文献

[1] 甘利俊一、『情報理論』、ちくま文庫、2011 年 4 月 10 日